RAIRO-Inf. Theor. Appl. **40** (2006) 53-74
DOI: 10.1051/ita:2005039

THEORIES OF ORDERS ON THE SET OF WORDS [*]

Dietrich Kuske[1]

Abstract. It is shown that small fragments of the first-order theory of the subword order, the (partial) lexicographic path ordering on words, the homomorphism preorder, and the infix order are undecidable. This is in contrast to the decidability of the monadic second-order theory of the prefix order [M.O. Rabin, *Trans. Amer. Math. Soc.*, 1969] and of the theory of the total lexicographic path ordering [P. Narendran and M. Rusinowitch, *Lect. Notes Artificial Intelligence*, 2000] and, in case of the subword and the lexicographic path order, improves upon a result by Comon & Treinen [H. Comon and R. Treinen, *Lect. Notes Comp. Sci.*, 1994]. Our proofs rely on the undecidability of the positive Σ_1-theory of $(\mathbb{N}, +, \cdot)$ [Y. Matiyasevich, *Hilbert's Tenth Problem*, 1993] and on Treinen's technique [R. Treinen, *J. Symbolic Comput.*, 1992] that allows to reduce Post's correspondence problem to logical theories.

Mathematics Subject Classification. 03D35.

1. Introduction

Depending on the context, the set of words carries different interesting partial orders. Seen as a free monoid, the divisor or *infix relation* is most natural; seen as nodes of the complete n-ary tree, one looks at the predecessor relation of the tree which coincides with the *prefix relation.* In combinatorics and the theory of well quasi orders, the *subword order* is central. In some cases, this order coincides with the *lexicographic path ordering* that is of outmost importance in string rewriting. The *homomorphism order* is closely related to the order of k-partitions over NP.

By Rabin's Theorem [23], the theory of the prefix order $(\Sigma^*, \leq)$ is decidable (this holds even for the monadic second order theory). Comon & Treinen [6] and Narendran & Rusinowitch [22] considered structures $(\Sigma^*, (p_a)_{a\in\Sigma}, \leq_{\text{lpo}})$ where p_a is a unary function that prefixes a word with the letter a and $\leq_{\text{lpo}}$ is the lexicographic path ordering [9]. In case this order is total, the theory of this structure is

[*] *These results were obtained while the author was affiliated with the Technische Universität Dresden.*

[1] Universität Leipzig, Germany; kuske@informatik.uni-leipzig.de

decidable [22]; the Σ_4-fragment is undecidable if the order is partial [6]. The latter undecidability holds for $(\Sigma^*, (p_a)_{a\in\Sigma}, \leq)$ as well where $\leq$ is the subword relation [6]. Here, we are mainly interested in the order relation alone, *i.e.*, in structures of the form $(\Sigma^*, \sqsubseteq)$ for some order relation $\sqsubseteq$. The existential theory turns out to be decidable in all cases since all finite partial orders can be embedded. For small alphabets, we show the full theory of the homomorphism preorder $(\Sigma^*, \lesssim)$ to be decidable. Apart from these results, we prove the undecidability of small fragments of the theory of all the abovementioned partial orders for two-elements alphabets.

We start with the subword order $\hookrightarrow$ where $u \hookrightarrow v$ iff u results from v by deleting some occurrences of letters. This relation has been studied extensively. It is a well order [15], the homotopy types of its intervals [11] and rational expressions for related Möbius-functions [2] have been calculated, it has been considered under counting aspects Chapter 6 of [19], in the context of Macaulay-posets [8, 18] and of formal languages [14], and it can be used to show decidability results on "lossy channel systems" [12] and asynchronously cellular automata [17]. We reduce the positive Σ_1-fragment of arithmetic to the Σ_3-theory of the subword order $(\{a, b\}^*, \hookrightarrow)$ which proves the undecidability of this fragment (Th. 2.3).

In rewriting theory, it is desirable to orient a set of equations as a first step towards a terminating and confluent rewrite system. A typical example, where this is not possible, is the set of equations defining commutative groups: the commutativity law cannot be ordered in any way. A possible solution is to consider ordered rewriting: $s \cdot t$ is rewritten into $t \cdot s$ only in case $s < t$ for some order relation $\leq$. This can be extended by allowing more involved properties than $s < t$, *e.g.*, arbitrary first-order formulas. To apply such a strategy in an automatic system, these properties have to be decidable. Refining the proof for the subword order, we show that the Σ_2-theory of $(\{a, b\}^*, p_a, p_b, \leq_{\mathrm{lpo}})$ is undecidable where $\leq_{\mathrm{lpo}}$ is a partial lexicographic path ordering (Th. 3.5). This improves upon a result by Comon & Treinen [6] stating this result for the Σ_4-theory in case the alphabet contains at least three letters. For the pure ordered structure $(\{a, b\}^*, \leq_{\mathrm{lpo}})$, we obtain the undecidability of the Σ_3-theory.

In [16], Kosub & Wagner study the structure of k-partitions over NP. Generalizing the classes of the Boolean Hierarchy over NP, they define classes of k-partitions and investigate their order structure which turns out to be rather complicated in general. Any such class is given by a $\{1, 2, \ldots, k\}$-labeled finite lattice. Their embedding conjecture states that the order on these classes is completely described by the homomorphism order between the defining labeled lattices. This conjecture is proved in some cases, in particular if the defining lattice is a chain, *i.e.*, a word over an k-elements alphabet. Therefore, we turn attention to the "homomorphism preorder" on words: consider a word of length n as a labeled linear order with n elements (the labels come from the alphabet). Then $u \lesssim v$ if there exists a homomorphism from the labeled linear order representing u into the labeled linear order representing v. Equivalently, $u \lesssim v$ if, after deleting repetitions in u, we obtain a subword of v. Interpreting the subword relation in the homomorphism preorder (of a larger alphabet), it follows that the Σ_5-theory of this preorder is undecidable for alphabets with at least four elements (Th. 4.3); the full theory is decidable for alphabets with at most two elements (*cf.* discussion after Th. 4.3).

Makanin [20] showed that the positive Σ_1-theory of the free monoid is decidable (the extension to the full Σ_1-theory can be found in [5]). The positive Σ_1-theory of the divisor or infix relation $\leq$ is easily reduced to the Σ_1-theory of the free monoid; hence it is decidable as well. On the other hand, the positive Π_2-theory of $(\Sigma^*, \cdot)$ was shown to be undecidable by Durnev [10], but there is no obvious reduction of this theory to the theory of the infix relation. To show the undecidability of the Σ_4-theory of $(\Sigma^*, \leq)$ (*cf.* Th. 5.6), we use the ideas developed by Treinen in [25]. There, he gives a general result stating that, provided formulas with certain properties exist in a given theory, the underlying theory is undecidable. The point is that the existence of these formulas allows to encode Post's correspondence problem in the theory. I decided not to start from his criteria, but to follow his proof in the concrete setting of this particular partial order. The reasons for this decision are two-fold: it should simplify the understanding since Post's correspondence problem is directly encoded into the theory, and it is not much longer than showing that our formulas have the properties required by Treinen.

In the final section, we extend our focus to finite Σ-labeled forests. The analogon of the subword relation is the embeddability: the forest s is smaller than the forest t if s can be embedded into t. We show that this order relation has an undecidable theory (Cor. 6.1).

I would like to thank Denis Lugiez for directing my attention to lexicographic path orderings and pointing me to the results of Comon & Treinen, Ralf Treinen for his pointer to Narendran & Rusinowitch's work, and Victor Selivanov for his continuous interest in the results to be reported in this article (although I could not solve his original question).

Notation. Concerning first order logic, we only recall that Σ_n stands for the set of formulas that are logically equivalent to some formula in prenex normal form with n blocks of quantifiers (starting with a block of existential quantifiers); Π_n is the set of negations of formulas in Σ_n. By $B\Sigma_n$, we denote the set of Boolean combinations of formulas from Σ_n. The Σ_n-theory of a structure $\mathcal{S}$ is the set of sentences from Σ_n that hold in $\mathcal{S}$.

2. The subword order

For two words u and v over some alphabet Σ, we write $u \hookrightarrow v$ if u is a subword of v, *i.e.*, if u results from v by deleting arbitrary many letters. Equivalently, $u \hookrightarrow v$ if u (seen as a Σ-labeled linear order) can be embedded into v.

Theorem 2.1 (Comon & Treinen [6]). *Let $a, b, \# \in \Sigma$ and let the function $p_\# : \Sigma^* \to \Sigma^*$ be defined by $w \mapsto \#w$. Then the Σ_4-theory of $\mathcal{S} = (\Sigma^*, p_\#, \hookrightarrow, a, b)$ is undecidable.*

More precisely, Comon & Treinen show that the free monoid $(\{a, b\}^*, \cdot, a, b)$ can be interpreted in $\mathcal{S}$. All the formulas in this interpretation are at most Σ_3. Since the positive $\forall\exists^3$-theory of the free monoid is undecidable [10], one obtains the undecidability of the Σ_4-theory of $\mathcal{S}$.

It is the aim of this section to sharpen the above result. In particular, we will show that the Σ_3-theory of $(\Sigma^*, \hookrightarrow)$ (which is a reduct of Comon & Treinen's structure) is undecidable as soon as Σ contains at least 2 elements.

But first, we prove that the existential theory of $(\Sigma^*, \hookrightarrow)$ is decidable.

Proposition 2.2. *Let Σ be some alphabet. Then the Σ_1-theory of $(\Sigma^*, \hookrightarrow)$ is decidable.*

Proof. Since $(\{a\}^*, \hookrightarrow) \cong (\mathbb{N}, \leq)$, this holds for $|\Sigma| = 1$. For larger alphabets, any finite partial order can be embedded into $(\Sigma^*, \hookrightarrow)$: Let $(\{p_1, p_2, \ldots, p_n\}, \preceq)$ be a finite partial order and choose $a, b, c \in \Sigma$ with $c \notin \{a, b\}$ (in a later proof, we will use that these three letters are mutually distinct). For $j = 1, 2, \ldots, n$, set $w_j = abx_1abx_2abx_3 \ldots x_na$ with $x_i = c$ iff $p_i \preceq p_j$, and $x_i = \varepsilon$ otherwise. For any of the words w_i, the maximal subword from $\{a, b\}^*$ is $(ab)^n a$. Hence we get $w_i \hookrightarrow w_j$ iff w_i results from w_j by the deletion of some occurrences of the letter c, hence iff $p_i \preceq p_j$. Hence the formula $\exists x_1 \exists x_2 \ldots \exists x_n : \psi$ where ψ is quantifier-free holds in $(\Sigma^*, \hookrightarrow)$ iff it holds in some finite partial order whose size can be restricted to n. Since there are only finitely many such orders, the result follows. □

We now turn to the announced undecidability proof. Let $\Sigma = \{a, b\}$ and $\mathcal{M} = (\Sigma^*, \hookrightarrow, \varepsilon, a, b, ab, ba, aa, bb, aba, bab)$. We will encode arithmetic by identifying a natural number n with the word a^n. Projections will be the main tool in this section. *E.g.*, the longest a-prefix of a word from a^*ba^* is the projection to a^* of its projection to a^*b. A word contains precisely k occurrences of the letter a iff its projection to a^* is a^k. Hence addition can be defined by "there is a word from a^*ba^* with longest a-prefix a^m and longest a-suffix a^n whose projection to a^* is a^{m+n}". To encode multiplication, we show that the binary relation "same length" can be defined. This allows to define the word b^m from a^m and then, from b^m and a^n, the word $(ba^n)^m$. Its projection to a^* is $a^{m \cdot n}$.

First, we have to show that some auxiliary relations are definable in $\mathcal{M}$. For a word $w \in \Sigma^*$, let $\pi_a(w)$ denote the projection of w to a^*, *i.e.*, the largest element of a^* that embeds into w; $\pi_b(w)$ is defined similarly.

(S1) A word $w \in \Sigma^*$ belongs to a^* iff it does not contain any occurrence of the letter b, *i.e.*, iff $b \not\hookrightarrow w$. Hence the sets a^* and b^* are Σ_0-definable in $\mathcal{M}$.

(S2) $\pi_a(w)$ is the maximal word $u \in a^*$ embedding into w, *i.e.*, $u = \pi_a(w)$ iff $u \in a^* \wedge u \hookrightarrow w \wedge \forall x((u \hookrightarrow x \hookrightarrow w \wedge x \in a^*) \Rightarrow (u = x))$. Hence the set of all pairs $(u, w) \in \Sigma^* \times \Sigma^*$ with $u = \pi_a(w)$ is Π_1-definable in $\mathcal{M}$. The projection to b^* can be handled similarly.

(S3) Let $x = a^m ba^n$ and $u \in a^*$. Then $u = a^m$ iff the following holds in $\mathcal{M}$:

$$\begin{array}{lll} & \forall y, z & ((b \hookrightarrow y \hookrightarrow x \wedge ba \not\hookrightarrow y \wedge b \not\hookrightarrow z \hookrightarrow y) \Rightarrow (z \hookrightarrow u)) \\ \wedge & \exists y & (b \hookrightarrow y \wedge ba \not\hookrightarrow y \wedge u \hookrightarrow y \hookrightarrow x). \end{array}$$

The condition in the first line expresses that $y \in a^*b^+$ is a subword of x (and therefore contained in a^*b) and $z \in a^*$ is a subword of y. Hence (with

$x = a^m ba^n$), the length of z is bounded by m. The conclusion says that u contains at least $|z|$ occurrences of a. Since z can take the value a^m, we have that u is at least a^m. The second line expresses that in particular $y = ub$ is a subword of x, *i.e.*, that u contains at most m occurrences of a. The formula above is $B\Sigma_1$.

Now we are in the position to show that $S = \{(a^m, a^n, a^{m+n}) \mid m, n \in \mathbb{N}\}$ is definable: $(u, v, w) \in S$ iff

$$\exists x : \left(\begin{array}{ll} & b \hookrightarrow x \wedge \neg(bb \hookrightarrow x) \wedge w = \pi_a(x) \\ \wedge & u \in a^* \wedge v \in a^* \\ \wedge & u \text{ is the maximal } a^*\text{-prefix of } x \\ \wedge & v \text{ is the maximal } a^*\text{-suffix of } x \end{array} \right) .$$

The first three conjuncts state that $x = a^m ba^n$ and $w = a^{m+n}$ for some $m, n \in \mathbb{N}$. Together with the second line, x and u as well as x and v satisfy the assumption in (S3). Hence the third and fourth lines ensure $u = a^m$ and $v = a^n$ (which is $B\Sigma_1$-expressible by (S3)). Thus, indeed, S is Σ_2-definable.

We define some more auxiliary relations

(S4) A word $w \in \Sigma^*$ belongs to $\Sigma^* b$ iff

$$\exists x (x \hookrightarrow w \wedge \forall y((y \in a^* \wedge y \hookrightarrow w) \Rightarrow (y \hookrightarrow x)) \wedge b \hookrightarrow x \wedge ba \not\hookrightarrow x).$$

By the first conjunct, x results from w by deleting some letters. The second conjunct ensures that all occurrences of a are still present, *i.e.*, that x results from w by deleting some occurrences of b. The final two conjuncts express that the last letter of x is b. Thus, indeed, this Σ_2-formula holds iff $w \in \Sigma^* b$. Similarly, the set $a\Sigma^*$ is Σ_2-definable.

(S5) $w \in \Sigma^* \setminus (\Sigma^* bb \Sigma^*)$ iff $\forall u(bb, \pi_a(w) \hookrightarrow u \hookrightarrow w \Rightarrow bab \hookrightarrow u)$: the condition expresses that u results from w by deleting some occurrences of b and that at least two occurrences of b are left. The conclusion requires that these two occurrences of b are separated by some a. To place this expression in Σ_2, we use the alternative formulation

$$\exists x (x = \pi_a(w) \wedge \forall u(bb, x \hookrightarrow u \hookrightarrow w \Rightarrow bab \hookrightarrow u)).$$

It is indeed Σ_2 since $x = \pi_a(w)$ is expressible in Π_1 by (S2).

(S6) The set $(ab)^+ = a\Sigma^* \cap \Sigma^* b \cap \Sigma^* \setminus (\Sigma^* aa \Sigma^* \cup \Sigma^* bb \Sigma^*)$ is Σ_2-definable by (S4) and (S5). The same applies to $(ab)^* = (ab)^+ \cup \{\varepsilon\}$.

(S7) A pair of words (u, v) belongs to $E = \{(a^n, b^n) \mid n \in \mathbb{N}\}$ iff there exists $x \in (ab)^*$ with $u = \pi_a(x)$ and $v = \pi_b(x)$; hence E is Σ_2-definable by (S2) and (S6).

(S8) Let $u = b^m$, $v = a^n$ and $z \in \Sigma^*$. Then $z = a^{n_0} ba^{n_1} ba^{n_2} \dots ba^{n_m}$ for some $n_i \le n$ iff

$$\varphi_{(S8)}(z) \equiv (u = \pi_b(z) \wedge \forall x((u \hookrightarrow x \hookrightarrow z \wedge aba \not\hookrightarrow x) \Rightarrow (\pi_a(x) \hookrightarrow v))).$$

The formula $u = \pi_b(z)$ forces z to be of the form $a^{n_0}ba^{n_1}ba^{n_2}\dots ba^{n_m}$ for some $n_i \in \mathbb{N}$. Let $x = b^i a^{n_i} b^{m-i}$ with $0 \leq i \leq m$. Then the premise of the above implication is satisfied. The conclusion ensures $a^{n_i} = \pi_a(x) \hookrightarrow v = a^n$, *i.e.*, $n_i \leq n$. If, conversely, z is of the desired form and x satisfies the premise of the implication, then $x = b^i a^k b^{m-i}$ for some $0 \leq i \leq m$. Since $x \hookrightarrow z$, we get $k \leq n_i \leq n$ and therefore $\pi_a(x) \hookrightarrow v$. We can express $\pi_a(x) \hookrightarrow v$ by the Π_1-formula $(\forall y(b \not\hookrightarrow y \hookrightarrow x \Rightarrow y \hookrightarrow v))$. Note that it occurs in a positive position under the universal quantifier $\forall x$, *i.e.*, the second conjunct of the above formula is in Π_1. By (S2), the same applies to the first conjunct. Thus, the formula $\varphi_{(S8)}$ is Π_1.

(S9) A triple of words (u, v, w) belongs to the set $\{(b^m, a^n, (ba^n)^m) \mid m, n \in \mathbb{N}\}$ iff $u \in b^*$, $v \in a^*$ and if w is the maximal element $z \in b\Sigma^*$ satisfying the formula $\varphi_{(S8)}(z)$ from (S8). This can be expressed as

$$a \not\hookrightarrow u \wedge b \not\hookrightarrow v \wedge w \in b\Sigma^* \wedge \forall z((z \in b\Sigma^* \wedge \varphi_{(S8)}(z)) \Rightarrow z \hookrightarrow w).$$

The formula $w \in b\Sigma^*$ is Σ_2 by (S4). The formula $z \in b\Sigma^* \wedge \varphi_{(S8)}(z)$ is Σ_2 by (S4) and (S8). Since it occurs at a negative position under the universal quantification $\forall z$, the last conjunct of the above formula is Π_2. Hence the whole formula above is $B\Sigma_2$.

Now we are in the position to show that $P = \{a^m, a^n, a^{m\cdot n} \mid m, n \in \mathbb{N}\}$ is definable as well: $(u, v, w) \in P$ iff

$$\exists x, y : \left(\begin{array}{ll} & (u, x) \in E \\ \wedge & (x, v, y) \in \{(b^m, a^n, (ba^n)^m) \mid m, n \in \mathbb{N}\} \\ \wedge & w = \pi_a(y) \end{array} \right).$$

The first conjunct under the existential quantifier states that $u = a^m$ and $x = b^m$ for some $m \in \mathbb{N}$. By the second conjunct, $v = a^n$ and $y = (ba^n)^m$ for some $n \in \mathbb{N}$. Hence, by the third conjunct, $w = a^{m\cdot n}$ as required. By (S7), the first conjunct is a Σ_2-formula, by (S9), the second one is in $B\Sigma_2$, and by (S2), the last one is in Π_1. Hence, altogether, this formula is in Σ_3.

We showed that a^*, S, and P can be defined in $\mathcal{M}$. Since $(a^*, S, P) \cong (\mathbb{N}, +, \cdot)$, this implies that the theory of $\mathcal{M}$ is undecidable [13].

Theorem 2.3. *The* Σ_3*-theory of* $(\{a, b\}^*, \hookrightarrow)$ *is undecidable.*

Proof. So far, we found a Σ_3-interpretation of $(\mathbb{N}, +, \cdot)$ in

$$\mathcal{M} = (\Sigma^*, \hookrightarrow, \varepsilon, a, b, ab, ba, aa, bb, aba, bab).$$

Since the positive Σ_1-theory of $(\mathbb{N}, +, \cdot)$ is undecidable by [21], the Σ_3-theory of $\mathcal{M}$ is undecidable. As a first step, we want to reduce the Σ_3-theory of $\mathcal{M}$ to that of $\mathcal{M}' = (\Sigma^*, \hookrightarrow, a, b, ab)$. For this, we show how to define ba etc. in $\mathcal{M}$.

- ε is the least word w of $\mathcal{M}'$, a property expressible by a Π_1-formula φ_ε.

- A word $w \in \Sigma^*$ equals ba iff it is not ab, and any of its proper subwords is a subword of a or of b. This is expressed by the following Π_1-formula:

$$\varphi_{ba}(w) \equiv a, b \hookrightarrow w \neq ab \wedge \forall x(x \hookrightarrow w \iff (x = w \vee x \hookrightarrow a \vee x \hookrightarrow b)).$$

- The word aa is the minimal element of $a^* \setminus \{\varepsilon, a\}$, *i.e.*, $w = aa$ iff

$$\varphi_{aa}(w) \equiv w \in a^* \wedge \forall x((x \in a^* \wedge \neg\varphi_\varepsilon(x) \wedge x \neq a) \Rightarrow (w \hookrightarrow x)).$$

  The Π_1-formula φ_{bb} is defined similarly.
- The word aba is the least word w embedding ab and ba, but not bb, *i.e.*, $w = aba$ iff $\varphi_{aba}(w) \equiv ab, ba \hookrightarrow w \wedge b^2 \not\hookrightarrow w \wedge \forall x((ab, ba \hookrightarrow x \wedge bb \not\hookrightarrow x) \Rightarrow (w \hookrightarrow x))$ which is a Π_1-formula. Note that this is not a formula in the language of $(\Sigma^*, \hookrightarrow, a, b, ab)$, but it mentions the additional constants ba and bb. The Π_1-formula φ_{bab} is defined similarly (using aa as an additional constant).

Now let φ be a Σ_3-sentence in the language of $\mathcal{M}$. Then $\mathcal{M} \models \varphi$ iff $\mathcal{M}'$ satisfies

$$\exists w_\varepsilon, w_{ba}, w_{aa}, w_{bb}, w_{aba}, w_{bab} : \begin{pmatrix} \varphi_\varepsilon(w_\varepsilon) \wedge \varphi_{ba}(w_{ba}) \wedge \varphi_{aa}(w_{aa}) \wedge \varphi_{bb}(w_{bb}) \\ \wedge\, \varphi'_{aba}(w_{aba}) \wedge \varphi'_{bab}(w_{bab}) \wedge \varphi' \end{pmatrix}.$$

Here φ' and φ'_w result from φ and φ_w, resp., by replacing ab by w_{ab} etc. Since all the conjuncts are at most Σ_3-formulas, the whole formula is Σ_3, *i.e.*, we reduced the undecidable Σ_3-theory of $\mathcal{M}$ to the Σ_3-theory of $\mathcal{M}'$. It remains to get rid of the constants a, b, and ab. Since the models $(\Sigma^*, \hookrightarrow, a, b, ab)$ and $(\Sigma^*, \hookrightarrow, b, a, ba)$ are isomorphic (replacing any occurrence of a in a word w by b and *vice versa* is such an isomorphism), we cannot define the elements a, b, and ab in $(\Sigma^*, \hookrightarrow)$, but we can define them "up to isomorphism":

For a Σ_3-sentence φ in the language of $\mathcal{M}'$, we consider the formula $\overline{\varphi}$

$$\exists w_a, w_b, w_{ab} : \begin{pmatrix} & w_a \not\hookrightarrow w_b \wedge w_b \not\hookrightarrow w_a \\ \wedge & \exists w_\varepsilon \forall x((x \hookrightarrow w_a) \Rightarrow (x = w_\varepsilon \vee x = w_a)) \\ \wedge & \exists w_\varepsilon \forall x((x \hookrightarrow w_b) \Rightarrow (x = w_\varepsilon \vee x = w_b)) \\ \wedge & \forall x(x \hookrightarrow w_{ab} \Leftrightarrow (x = w_{ab} \vee x \hookrightarrow w_a \vee x \hookrightarrow w_b)) \\ \wedge & \varphi' \end{pmatrix}$$

where φ' results from φ by replacing any occurrence of a by w_a, b by w_b and ab by w_{ab}.

The first three lines ensure that w_a and w_b are two distinct elements of Σ. The fourth line holds iff $w_{ab} \in \{w_a w_b, w_b w_a\}$. Hence we get $(\Sigma^*, \hookrightarrow, a, b, ab) \cong (\Sigma^*, \hookrightarrow, w_a, w_b, w_{ab})$ which implies $\mathcal{M}' \models \varphi \iff (\Sigma^*, \hookrightarrow) \models \overline{\varphi}$. □

Corollary 2.4. *Let $\Sigma = \{a_1, a_2, \ldots, a_n\}$ be an alphabet with at least two elements. Then the Σ_3-theory of the subword order $(\Sigma^*, \hookrightarrow)$ is undecidable.*

Proof. Let φ be a Σ_3-sentence in the language of $(\{a,b\}^*, \hookrightarrow)$ and consider the following formula $\overline{\varphi}$:

$$\exists (w_i)_{i\leq n} \left(\begin{array}{ll} & \bigwedge_{1\leq i<j\leq n} w_i \not\hookrightarrow w_j \wedge w_j \not\hookrightarrow w_i \\ \wedge & \exists w_\varepsilon \bigwedge_{i\leq n} \forall x (x \hookrightarrow w_i \iff (w = w_\varepsilon \vee w = w_i)) \\ \wedge & \varphi' \end{array} \right) .$$

Here, the formula φ' results from φ by simultaneous replacement of $\exists x\psi$ by $\exists x(\bigwedge_{3\leq i\leq n} \neg w_i \hookrightarrow x \wedge \psi)$ and of $\forall x\psi$ by $\forall x(\bigwedge_{3\leq i\leq n} \neg w_i \hookrightarrow x \Rightarrow \psi)$. Then $(\Sigma^*, \hookrightarrow) \models \overline{\varphi}$ iff the restriction of φ to $\{w_1, w_2\}^*$ holds in $(\Sigma^*, \hookrightarrow)$. Since w_1 and w_2 are distinct letters from Σ, we obtain $(\{a,b\}^*, \hookrightarrow) \models \varphi$ iff $(\Sigma^*, \hookrightarrow) \models \overline{\varphi}$. □

3. The lexicographic path order

Let Γ be some finite functional signature (*i.e.*, a finite set of function and constant symbols with associated arity) and $T(\Gamma)$ the associated set of Γ-terms. Further, with any n-ary function symbol $f \in \Gamma$, one associates an n-ary operation on $T(\Gamma)$ (also denoted f) with $(t_1, \dots, t_n) \mapsto f(t_1, \dots, t_n)$. The signature Γ is *unary* if it consists of unary and constant function symbols, only. Any partial order $\leq$ on Γ defines a lexicographic path ordering $\leq_{\text{lpo}}$ [9] on the set of Γ-terms $T(\Gamma)$ (see also [1]). The actual definition of lexicographic path orderings is nontrivial and, for our technical arguments, only the following observation is of importance (it is folklore in the rewriting community):

Observation. Let Σ be some alphabet and set $\Gamma_\Sigma = \Sigma \cup \{\perp\}$ where the letters from Σ are unary and $\perp$ is the only constant symbol. On Γ_Σ, consider the precedence $\leq$ with $\perp < a$ for all $a \in \Sigma$ and no further comparabilities hold. Then the structures $(\Sigma^*, (p_a)_{a\in\Sigma}, \hookrightarrow)$ (*cf.* Th. 2.1) and $(T(\Gamma_\Sigma), \Gamma_\Sigma, \leq_{\text{lpo}})$ are isomorphic.

For general signatures, the following is known.

Theorem 3.1 (Comon & Treinen [7]). *Let Γ be a signature containing a constant, an at least unary and an at least binary symbol. Then there exists a total precedence $\leq$ on Γ such that the Σ_2-theory of the structure $(T(\Gamma), \Gamma, \leq_{\text{lpo}})$ is undecidable.*

Any signature has to contain a constant for otherwise there were no terms. Most likely, one can discard the unary symbol from the signature and still get the result (*cf.* discussion at the end of [7]). But the proof makes crucial use of the binary symbol. A signature is *unary* if it consists of unary and constant symbols, only. Narendran & Rusinowitch showed that the binary symbol is crucial not only in the proof by Comon & Treinen, but for the result to hold:

Theorem 3.2 (Narendran & Rusinowitch [22]). *Let Γ be a unary signature with total precedence $\leq$. Then the theory of $(T(\Gamma), \Gamma, \leq_{\text{lpo}})$ is decidable.*

By Theorem 3.1, the signature has to be unary for this result to hold. From Theorem 2.1, one also gets that totality of the precedence relation is necessary as we indicate now.

The isomorphism of the structures $(\Sigma^*, (p_a)_{a\in\Sigma}, \hookrightarrow)$ and $(T(\Gamma_\Sigma), \Gamma_\Sigma, \leq_{\mathrm{lpo}})$ and Theorem 2.1 imply

Corollary 3.3 (Comon & Treinen [6])**.** *Let Γ be a unary signature with at least three function symbols and one constant symbol. Then there exists a partial precedence $\leq$ on Γ such that the Σ_4-theory of $(T(\Gamma), \Gamma, \leq_{\mathrm{lpo}})$ is undecidable.*

Our Corollary 2.4 allows to similarly derive the slightly stronger result

Corollary 3.4. *Let Γ be a unary signature with at least two function symbols and one constant symbol. Then there exists a partial precedence $\leq$ on Γ such that the Σ_3-theory of $(T(\Gamma), \leq_{\mathrm{lpo}})$ is undecidable.*

Note that, differently from Corollary 3.3, this result speaks about the reduct $(T(\Gamma), \leq_{\mathrm{lpo}})$ of the structure $(T(\Gamma), \Gamma, \leq_{\mathrm{lpo}})$ that Comon & Treinen were interested in. For arbitrary signatures and *total* precedence, the theory of this reduct is decidable since the partial order $(T(\Gamma), \leq_{\mathrm{lpo}})$ is an ordinal in that case [4]. We now show that one quantifier alternation and two function symbols suffice for the undecidability in Corollary 3.3 (it is not clear whether Cor. 3.4 holds for Σ_2). For this, we refine our interpretation of $(\mathbb{N}, +, \cdot)$ in $(\{a,b\}^*, \hookrightarrow, \varepsilon, a, b, ab, ba, aa, bb, aba, bab)$ from the previous section. Now, we interpret $(\mathbb{N}, +, \cdot)$ in

$$\mathcal{M} = (\{a,b\}^*, p_a, p_b, \hookrightarrow, \varepsilon, a, b, ab, ba, aa, bb, aba, bab),$$

i.e., in addition we can prefix a word with a given letter. This allows to express $(u, v, w) \in \{(b^m, a^n, (ba^n)^m) \mid m, n \in \mathbb{N}\}$ (*cf.* (S9)) using the Σ_2-formula

$$\begin{array}{ll} & a \not\hookrightarrow u \wedge b \not\hookrightarrow v \wedge \pi_b(w) = u \wedge \exists x(w = bx) \\ \wedge & \forall x((u \hookrightarrow x \hookrightarrow w \wedge aba \not\hookrightarrow x) \Rightarrow (\pi_a(x) \hookrightarrow v)) \\ \wedge & \forall x((u, v \hookrightarrow bx \wedge aba, av, bu \not\hookrightarrow bx) \Rightarrow (bx \hookrightarrow w)) \\ \wedge & \exists x(x \in (ba)^* \wedge u \hookrightarrow x \hookrightarrow w). \end{array}$$

The first two conjuncts ensure $u = b^m$ and $v = a^n$ for some $m, n \in \mathbb{N}$. Since $\pi_b(w) = u$, the word w contains precisely m occurrences of b and starts with b (since $\exists x(w = bx)$). The second line ensures that any a-block in w has size at most n (*cf.* discussion in (S8)). Now consider the third line. A word x satisfies the premise iff

- bx contains the same number of occurrences of b as u and therefore w does (since $u \hookrightarrow bx \wedge bu \not\hookrightarrow bx$);
- bx contains the same number of occurrences of a as v does (since $v \hookrightarrow bx \wedge av \not\hookrightarrow bx$);
- bx contains precisely one a-block (since $aba \not\hookrightarrow bx$).

Thus, x satisfies the premise iff $bx = b^i a^n b^{m-i}$ for some $1 \leq i \leq m$. Hence the third line ensures that any such word embeds into w, *i.e.*, any nonempty a-block in w has size at least n. The last line is meant to ensure that there is a nonempty a-block between any two consecutive bs and at the end of w.

Using this formula instead of the formula from (S9) yields a Σ_2-description of the relation $P = \{(a^m, a^n, a^{m\cdot n} \mid m, n \in \mathbb{N}\}$. Hence we have a Σ_2-interpretation of $(\mathbb{N}, +, \cdot)$ in the structure $\mathcal{M}$. Since the positive Σ_1-theory of $(\mathbb{N}, +, \cdot)$ is undecidable, the Σ_2-theory of $\mathcal{M}$ is undecidable. Let φ be a Σ_2-sentence in the language of $\mathcal{M}$ and replace any occurrence of, *e.g.*, ab in φ by $p_a p_b(\varepsilon)$. This yields a Σ_2-sentence φ' in the language of the structure $\mathcal{M}' = (\{a,b\}^*, p_a, p_b, \hookrightarrow, \varepsilon)$ that holds in this latter structure iff $\mathcal{M} \models \varphi$. Finally, consider $\psi = \exists w_\varepsilon(\forall w : (w_\varepsilon \hookrightarrow w) \wedge \varphi')$ where φ' results from φ by replacing ε by w_ε. Then we have $\mathcal{M} \models \varphi$ iff $(\{a,b\}^*, p_a, p_b, \hookrightarrow) \models \psi$. Since ψ is a Σ_2-sentence, considerations as in the proof of Corollary 2.4 yield

Theorem 3.5. *Let Γ be some unary signature with at least two function symbols and one constant symbol. Then there exists a partial precedence $\leq$ on Γ such that the Σ_2-theory of $(T(\Gamma), \Gamma, \leq_{\text{lpo}})$ is undecidable.*

4. The homomorphism preorder of words

Let $u = u_1 u_2 \dots u_n$ be some word of length n. With u, we associate the Σ-labeled linear order $\underline{u} = (\{1, 2, \dots, n\}, \leq, \lambda)$ with $\lambda(i) = u_i$. For two words u and v, let $u \lesssim v$ denote the existence of a homomorphism from $\underline{u}$ into $\underline{v}$ (*i.e.*, an order preserving function f from $\{1, 2, \dots, |u|\}$ to $\{1, 2, \dots, |v|\}$ with $u_i = v_{f(i)}$). For instance, $u \hookrightarrow v$ implies $u \lesssim v$ for any two words u and v. But also $aab \lesssim ab$ witnessed by the mapping f with $f(1) = f(2) = 1$ and $f(3) = 2$. The relation $\lesssim$ is a preorder that, to the knowledge of the author, appeared for the first time in [16] in the context of complexity theoretic considerations. Let $\sim \, = \, \lesssim \cap \gtrsim$ be the associated equivalence relation. Then $\lesssim/\sim$ is a partial order on $\Sigma^*/\sim$. The structure $(\Sigma^*/\sim, \lesssim/\sim)$ is Σ_0-interpretable in $(\Sigma^*, \lesssim)$. Showing that the Σ_5-theory of the former is undecidable therefore implies that the Σ_5-theory of the latter is undecidable.

Each $\sim$-equivalence class contains precisely one *repetition-free word* (*i.e.*, a word not containing aa as an infix for any letter a). For repetition-free words u and v, we have $u \lesssim v$ iff $u \hookrightarrow v$. Thus, the structure $(\Sigma^*/\sim, \lesssim/\sim)$ is isomorphic to the structure $(\text{RF}, \hookrightarrow)$ where RF denotes the set of repetition-free words. We first prove some decidability results:

Theorem 4.1. *If the alphabet Σ contains at most two letters, then the theory of $(\Sigma^*, \lesssim)$ is decidable.*

Proof. For one letter, we have only two repetition-free words (ε and a); hence the theory is decidable in this case. For two letters, the only repetition-free words are the elements of the set $\{a, \varepsilon\}(ba)^*\{b, \varepsilon\}$. Its nonempty elements can naturally be identified with the elements of the set $\{a, b\} \times \mathbb{N}$ by $w \mapsto$ (first letter of $w, |w|$). Then $u \hookrightarrow v$ iff they agree on their first letter and $|u| \leq |v|$, or they don't and $|u| < |v|$. Hence for two letters, the theory of $(\text{RF}, \hookrightarrow)$ can be interpreted in the theory of $(\mathbb{N}, \leq)$ which is decidable. □

The case of three letter alphabets remains open, later we will show the undecidability for four letters.

Proposition 4.2. *Let Σ be some alphabet. Then the Σ_1-theory of $(\Sigma^*, \lesssim)$ is decidable.*

Proof. By the previous theorem, we can assume the existence of three distinct letters a, b, c in Σ. Now the proof of Proposition 2.2 shows that the Σ_1-theory of $(\mathrm{RF}, \hookrightarrow)$ is decidable (since the embedding constructed there uses repetition-free words, only). Note that any $\sim$-equivalence class in Σ^* is infinite. Hence the Σ_1-theory of $(\Sigma^*, \lesssim)$ can be interpreted in that of its quotient $(\mathrm{RF}, \hookrightarrow)$. This ensures the result. □

To get the undecidability result, we will interpret $(\{a,b\}^*, \hookrightarrow)$ in $(\mathrm{RF}, \hookrightarrow)$. For this to work, two letters do not suffice by Theorem 4.1: let $\Sigma = \{a,b,c,d\}$ and $\Gamma = \{a,b\}$. Then $X = (c\{a,b\})^*$ is a set of repetition-free words over Σ. Let f be the monoid homomorphism from Γ^* to Σ^* with $a \mapsto ca$ and $b \mapsto cb$. Then f is injective and onto X, hence it witnesses $(X, \hookrightarrow) \cong (\Gamma^*, \hookrightarrow)$. It therefore remains to show that X is definable in $(\mathrm{RF}, \hookrightarrow)$. We will use the fact that a word $w \in \{a,b,c\}^*$ belongs to X iff, for any factorization $w = x_1x_2$ of w into two words, the last letter of x_1 or the first letter of x_2 equals c.

Similarly to the preceding sections, we work first in an extended model, namely in

$$\mathcal{M} = (\mathrm{RF}, \hookrightarrow, a, b, c, d, da, ad, db, bd, dc, cd, dcd).$$

(H1) Let $w \in \mathrm{RF}$. Then w belongs to $\{a,b,c\}^*$, to $\{a,b,c\}^*d$, or to $\{a,b,c\}^*dc$, resp., iff $d \not\hookrightarrow w$, iff $d \hookrightarrow w$ and $da, db, dc \not\hookrightarrow w$, or iff $dc \hookrightarrow w$, but $dcd, da, db \not\hookrightarrow w$, resp. Hence all these sets are Σ_0-definable.

(H2) Let $w, x \in \{a,b,c\}^* \cap \mathrm{RF}$. Then $x \hookrightarrow wdc$ iff the following Π_1-formula holds

$$\forall y((y \in \{a,b,c\}^*dc \wedge w \hookrightarrow y) \Rightarrow x \hookrightarrow y).$$

First suppose that this formula holds. Note that $y = wdc$ satisfies the premise of the implication. Thus, $x \hookrightarrow wdc$ follows. So suppose $x \hookrightarrow wdc$ and let $y \in \{a,b,c\}^*dc$ with $w \hookrightarrow y$. If we have $x \hookrightarrow w$, then $x \hookrightarrow w \hookrightarrow y$ ensures $x \hookrightarrow y$. So suppose $x \not\hookrightarrow w$. Since $x \hookrightarrow wdc$ and $x \in \{a,b,c\}^* \cap \mathrm{RF}$, there is a word x' with $x = x'c$ such that the last letter of x' is not c. Hence $x' \hookrightarrow w \hookrightarrow y = y'dc$ implies $x' \hookrightarrow y'$ since the last letter of x' is neither d nor c. Hence $x = x'c \hookrightarrow y'dc = y$. This finishes the proof.

(H3) Let $w \in \{a,b,c\}^*$. We show that c is the last letter of w iff

$$\forall x((x \in \{a,b,c\}^* \wedge x \hookrightarrow wdc) \Rightarrow x \hookrightarrow w).$$

(The formula expresses that w is the largest word from $\{a,b,c\}^* \cap \mathrm{RF}$ embedding into wdc.) If $w = w'a$ or $w = w'b$, then $w \hookrightarrow wc \hookrightarrow wdc$, *i.e.*, we showed "$\Leftarrow$". Conversely, suppose $w = w'c$ and $x \in \mathrm{RF} \cap \{a,b,c\}$ with $x \hookrightarrow wdc$. If the last letter of x is not c, then $x \hookrightarrow w$. So let $x = x'c$ for some word x'. Since x is repetition-free, the last letter of x' belongs to $\{a, b\}$. Hence $x'c = x \hookrightarrow wdc = w'cdc$ implies $x' \hookrightarrow w'$ and therefore $x \hookrightarrow w$. Thus, indeed, w is maximal in $\mathrm{RF} \cap \{a, b, c\}^*$ with $w \hookrightarrow wdc$.

Since the Π_1-subformula $x \hookrightarrow wdc$ appears in a negative position under the universal quantification $\forall x$, this formula is Π_2.

Analogously, we can express by a Π_2-formula that $w \in c\{a, b, c\}^* \cap \mathrm{RF}$. For this, we refer to the set $\{x \in \mathrm{RF} \mid x \hookrightarrow cdw\}$ (*cf.* (H3)) and therefore to the set $cd\{a, b, c\}^*$ (*cf.* (H2)). To define this latter one, we use the constants ad, bd, and cd analogously to (H1).

Our next aim is to express that $w = x_1x_2$.

(H4) Let $w \in \{a, b, c\}^* \cap \mathrm{RF}$. Then $x \in \mathrm{RF}$ satisfies the following Π_1-formula $\varphi_0(w, x)$

$$d \hookrightarrow x \wedge w \hookrightarrow x \wedge \forall x'((d \hookrightarrow x' \wedge w \hookrightarrow x' \hookrightarrow x) \Rightarrow x = x')$$

iff there are words $x_1, x_2 \in \mathrm{RF}$ with $w = x_1x_2$ and $x = x_1dx_2$.

(H5) Let $x = x_1dx_2$ with $x_1x_2 \in \{a, b, c\}^* \cap \mathrm{RF}$. Then $z = x_1$ iff the following Σ_2-formula $\varphi_1(x, z)$ is satisfied:

$$z \in \{a, b, c\}^* \wedge \exists y \left(\begin{array}{ll} & y \in \{a, b, c\}^*d \wedge y \hookrightarrow x \\ \wedge & \forall y'((y' \in \{a, b, c\}^*d \wedge y' \hookrightarrow x) \Rightarrow y' \hookrightarrow y) \\ \wedge & z \hookrightarrow y \\ \wedge & \forall z'((z' \in \{a, b, c\}^* \wedge z' \hookrightarrow y) \Rightarrow (z' \hookrightarrow z)) \end{array} \right) .$$

In the first line, $y \hookrightarrow x$ implies $y \hookrightarrow x_1d$ since x contains only one d and the last letter of y equals d. The second line therefore expresses that y is maximal with this property, *i.e.*, $y = x_1d$. Since z does not contain any d, the third line says $z \hookrightarrow x_1$. The last line therefore says that any d-free subword z' of x_1 is a subword of z, *i.e.*, $x_1 \hookrightarrow z$.

Let $\varphi_2(x, z)$ be the analogous Σ_2-formula expressing $z = x_2$.

Now we claim that a word $w \in \mathrm{RF} \cap \{a, b, c\}^*$ belongs to $X = (c\{a, b\})^*$ iff

$$\begin{array}{ll} & \forall x, z_1, z_2 \left(\begin{array}{l} (\varphi_0(w, x) \wedge \varphi_1(x, z_1) \wedge \varphi_2(x, z_2)) \\ \qquad \Rightarrow (z_1 \in \{a, b, c\}^*c \vee z_2 \in c\{a, b, c\}^* \vee z_1 = w) \end{array} \right) \\ & \wedge\ w \notin \{a, b, c\}^*c \\ \vee & (a \not\hookrightarrow w \wedge b \not\hookrightarrow w \wedge c \not\hookrightarrow w). \end{array}$$

Note that the last line is satisfied iff $w = \varepsilon$. So, from now on, we consider only nonempty words w. The premise in the first line is equivalent to $z_1z_2 = w$ by (H4) and (H5), *i.e.*, the universal quantification ranges over all factorizations of w into two words z_1 and z_2. First suppose that this formula holds. Let $w = z_1z_2$ be a factorization in two nonempty words z_1 and z_2. Then, by the conclusion in the first line, z_1 ends or z_2 begins with c. Hence, every other letter in w equals c. If z_1 in this factorization is empty, then $z_2 \in c\{a, b, c\}^*$ since $z_1 \neq w$. Hence $w = z_2$ starts with c. Thus, we showed that any odd letter of w equals c. Since w does not end with c and is repetition-free, we showed $w \in X$. Conversely, suppose

$w \in X \setminus \{\varepsilon\}$. Let $w = z_1 z_2$ be any factorization. If z_1 is empty, we obtain $z_2 = w$ which begins with c. If both, z_1 and z_2 are nonempty, then either z_1 ends or z_2 starts with c. If z_2 is empty, then $z_1 = w$. Thus, we showed the implication in the first line.

The formulas in the premise are from Σ_2 by (H4) and (H5) and the formulas in the conclusion in Π_2 by (H3). Since the implication appears under the universal quantification $\forall x, z_1, z_2$, the first conjunct is a Π_2-formula. Note that the second conjunct $w \notin \{a,b,c\}^* c$ is equivalent to $w \in \{a,b,c\}^* a \vee w \in \{a,b,c\} b$, *i.e.*, it is Π_2 as well. Thus, the set X is Π_2-definable in the structure $\mathcal{M}$. In other words, we found a Π_2-interpretation of $(\Gamma^*, \hookrightarrow)$ in $\mathcal{M}$. Since the Σ_3-theory of the former is undecidable, the Σ_5-theory of $\mathcal{M}$ is undecidable.

We next define the constants ad, db, bd, dc, cd, and dcd in the structure $\mathcal{M}' = (\mathrm{RF}, \hookrightarrow, a, b, c, d, da)$:

(H6) A word w is of length at least n iff there are mutually distinct $x_0 \hookrightarrow x_1 \hookrightarrow \dots x_{n-1} \hookrightarrow w$. Hence the set of words of length n is $B\Sigma_1$-definable in $\mathcal{M}'$.

(H7) $w = db$ iff $|w| = 2$, $d, b \hookrightarrow w$ and there are two distinct words $x \in \mathrm{RF}$ of length 3 with $da, w \hookrightarrow x$ (the other word v of length 2 with $d, b \hookrightarrow v$ is $v = bd$, but then bda is the only word from RF of length 3 embedding da and v). This formula is Σ_2. The word dc can be defined similarly.

(H8) $w = ad$ iff $w \neq da$, $a, d \hookrightarrow ad$, and $|w| = 2$. The words bd and cd are defined similarly by $B\Sigma_1$-formulas. (Note that these formulas use the constants db and dc which are not in the language of the structure $\mathcal{M}'$.)

(H9) $w = dcd$ iff $|w| = 3$, $dc, cd \hookrightarrow dcd$ and there is a word v of length 4 with $w, ad, da \hookrightarrow v$. (There is another word w' of length 3 with $dc, cd \hookrightarrow w'$, namely cdc. But for this one, we cannot construct v' of length 4 with $cdc, ad, da \hookrightarrow v'$.)

Thus, similarly to the first part of the proof of Theorem 2.3, the Σ_5-theory of $\mathcal{M}$ can be reduced to the Σ_5-theory of $\mathcal{M}'$. Considerations similar to those in the second part of the proof of Theorem 2.3 allow to interpret the Σ_5-theory of $\mathcal{M}$ in the Σ_5-theory of $(\mathrm{RF}, \hookrightarrow)$. Hence, we obtain

Theorem 4.3. *Let Σ be an alphabet with at least four elements. Then the Σ_5-theory of $(\mathrm{RF}, \hookrightarrow)$ and therefore that of $(\Sigma^*, \lesssim)$ is undecidable.*

Proof. Above, we explained how to prove the result about $(\mathrm{RF}, \hookrightarrow)$ if Σ contains precisely four letters. The proof of Corollary 2.4 yields the undecidability for larger alphabets. The model $(\mathrm{RF}, \hookrightarrow)$ is always isomorphic to $(\Sigma^*/{\sim}, \lesssim/{\sim})$. Since $\sim$ is Σ_0-definable in $(\Sigma^*, \lesssim)$, we can Σ_0-interpret $(\Sigma^*/{\sim}, \lesssim/{\sim})$ in $(\Sigma^*, \lesssim)$. Hence the Σ_5-theory of this latter structure is undecidable. □

5. The infix order

For two words u and v over Σ, we write $u \leq v$ if u is an infix of v, *i.e.*, if there exist words x, y such that $v = xuy$. In this section, we consider the theory of $(\Sigma^*, \leq)$.

Note that in the partial order $(\Sigma^*, \leq)$, any word dominates only finitely many other words (*i.e.*, any set $\{x \in \Sigma^* \mid x \leq w\}$ is finite). In the spirit of Propositions 2.2 and 4.2, we first show that any partial order with this property embeds into the infix order:

Lemma 5.1. *Let Σ be some alphabet and $a, b \in \Sigma$. Let $(P, \preceq)$ be some at most countable partial order such that any set $\{x \in P \mid x \preceq y\}$ is finite. Then $(P, \preceq)$ can be embedded into $(\Sigma^*, \leq)$.*

Proof. Without loss of generality, we assume $P \subseteq \mathbb{N} \setminus \{0\}$. By induction on the size $k(y)$ of the set $\{x \in P \mid x \prec y\}$, we define words w_y as follows: if $k(y) = 0$, set $w_y = ab^y a$. Otherwise, choose any enumeration $x_1, x_2, \dots, x_n$ of the set $\{x \in P \mid x \prec y\}$ and define $w_y = ab^y a\, w_{x_1}\, w_{x_2} \dots w_{x_n}$. Then $y \mapsto w_y$ is an order embedding of $(P, \preceq)$ into $(\Sigma^*, \leq)$. □

Proposition 5.2. *Let Σ be some alphabet. Then the Σ_1-theory of $(\Sigma^*, \leq)$ is decidable. If Σ is a singleton, then the full theory of $(\Sigma^*, \leq)$ is decidable.*

Proof. If Σ is a singleton, $(\Sigma^*, \leq)$ is isomorphic to $(\mathbb{N}, \leq)$ whose theory is decidable. If Σ contains at least two elements we can proceed as in Proposition 2.2 using Lemma 5.1. □

Our undecidability proof makes use of the following words over $\Sigma = \{\alpha, \beta\}$:

$$a = \alpha^6\beta \qquad b = \alpha^5\beta^2 \qquad \# = \alpha^4\beta^3 \qquad \bot = \alpha^3\beta^4 \qquad \#' = \alpha^2\beta^5 \qquad \bot' = \alpha\beta^6.$$

In particular, we will consider words over $\Gamma = \{a, b, \#, \bot, \#', \bot'\}$ and use them to encode the solvability of an instance of Post's correspondence problem. The encoding is based on ideas developed by Treinen [25]. To make the presentation self-contained, we only use his ideas, but do not refer to his results explicitly. The alphabet $\{a, b\}$ will be the alphabet of the instance, and the alphabet $\{\#, \bot, \#', \bot'\}$ will be used to encode the necessary operations.

We extend the structure $(\Sigma^*, \leq)$ by constants for all nonempty words of length at most 14: $\mathcal{M} = (\Sigma^*, \leq, (w)_{w \in \Sigma^+, |w| \leq 14})$. A word u is an *upper neighbor* of v iff $u \leq v$, $u \neq v$ and there is no word properly between these two.

Lemma 5.3. *A word $w \in \Sigma^*$ of length at least 14 belongs to Γ^* iff*

- *none of the words α^7, β^7, αu, $u\beta$ and uv with $u \in \Gamma$ and $v \in \Sigma^7 \setminus \Gamma$ is an infix of w;*
- *there are words $w' \in \Sigma^*$ and $u \in \Gamma$ such that w' is an upper neighbor of w in $(\Sigma^*, \leq)$ and $\alpha u \leq w'$; and*
- *there are words $w'' \in \Sigma^*$ and $v \in \Gamma$ such that w'' is an upper neighbor of w in $(\Sigma^*, \leq)$ and $v\beta \leq w''$.*

Hence the set Γ^* is Σ_2-definable in $\mathcal{M}$.

Proof. Suppose $u_i \in \Gamma$ for $1 \leq i \leq n$ ($n \geq 2$) and $w = u_1 u_2 \dots u_n$. Since all the words u_i have length 7 and belong to $\alpha^+\beta^+$, the word w does not contain α^7 or β^7 as an infix. Now let $u \in \Gamma$ and $x, y \in \Sigma^*$ with $xuy = w$. Then there is an index i

such that $x = u_1u_2 \dots u_{i-1}$, $u_i = u$, and $y = u_{i+1}u_{i+2} \dots u_n$. Hence the last letter of x is β (*i.e.*, αu is no infix of w) and the first letter of y is α (*i.e.*, $u\beta$ is no infix of w). Since the prefix of y of length 7 belongs to Γ, no word uv with $v \in \Sigma^7 \setminus \Gamma$ is an infix of w. Now let $w' = \alpha w$ and $u = u_1$. Then w' is an upper neighbor of w, $u \in \Gamma$, and αu is an infix of w'. The existence of w'' is shown similarly.

Conversely, let $w, w', w'' \in \Sigma^*$ and $u, v \in \Gamma$ satisfy all the requirements given in the lemma. Since $\alpha u \leq w'$ and $\alpha u \not\leq w$, the word u is a prefix of w. Similarly, v is a suffix of w. Let $u_i \in \Sigma^7$ and $x \in \Sigma^*$ with $|x| < 7$ and $w = u_1u_2 \dots u_n x$. We already showed $u_1 \in \Gamma$. Inductively, suppose $u_i \in \Gamma$ for some $i < n$. Then u_iu_{i+1} is an infix of w implying $u_{i+1} \in \Gamma$. In particular, the last letter of $u_n = \alpha^a\beta^{7-a}$ is β. Hence βx is a suffix of w of length at most 7 and therefore of $v \in \Gamma$. This implies $x = \beta^b$ for some b. Hence $\alpha^a\beta^{7+b-a}$ is a suffix of w. Since no $u\beta$ for $u \in \Gamma$ is an infix of w, we obtain $7 + b = a + 7 + b - a \leq 7$, *i.e.*, x is the empty word. Hence $w \in \Gamma^*$.

The first statement is quantifier-free. A quantifier alternation is needed in the second and the third statement in order to express that there are upper neighbors w' and w'' of w:

$$\exists w' \left(\begin{array}{ll} & w < w' \wedge \forall x(w < x \leq w' \Rightarrow x = w') \\ \wedge & \bigvee_{u\in\Gamma} \alpha u \leq w' \end{array} \right). \qquad \square$$

(I1) For $u, v \in \Gamma^*$, let neighbor(u, v) denote that there is $c \in \Gamma$ with $v \in \{cu, uc\}$; it expresses that v is an upper neighbor of u in $(\Gamma^*, \leq)$. Then neighbor(u, v) holds iff

$$\mathcal{M} \models u < v \wedge \neg\exists x_0, x_1, \dots, x_7, x_8 (u = x_0 \wedge \bigwedge_{1 \leq i \leq 8} x_{i-1} < x_i \wedge x_8 \leq v),$$

i.e., iff u is a proper infix of v and the length difference is at most 7 (since we assumed $u, v \in \Gamma^*$ from the very beginning, the length difference equals 7 in this case). Thus, the relation neighbor $\subseteq \Gamma^* \times \Gamma^*$ is Π_1-definable in $(\mathcal{M}, \Gamma^*)$ and therefore Σ_2-definable in $\mathcal{M}$ by Lemma 5.3.

(I2) Let $M \subseteq \Gamma$. Then a word $w \in \Gamma^*$ belongs to M^* iff $c \not\leq w$ for all $c \in \Gamma \setminus M$. Thus, any such set is Σ_0-definable in $(\mathcal{M}, \Gamma^*)$ by the formula

$$w \in \Gamma^* \wedge \bigwedge_{c \in \Gamma \setminus M} c \not\leq w$$

and therefore Σ_2-definable in $\mathcal{M}$.

(I3) A word $w \in \Sigma^*$ belongs to $\Sigma^*\# \setminus \Sigma^*\#\Sigma^+$ iff it contains #, but not $\#\gamma$ for any $\gamma \in \Sigma$. Hence this set is quantifier free definable in $\mathcal{M}$. Similarly, the set $\#\Sigma^* \setminus \Sigma^+\#\Sigma^*$ is Σ_0-definable using the constants $\alpha\#$ and $\beta\#$ from the signature of $\mathcal{M}$.

(I4) A word $w \in \Sigma^*$ belongs to $\#\{a, b\}^*\#$ iff it is an element of $\{a, b, \#\}^*$ and its two maximal proper infixes (within Γ^*) x_1 and x_2 satisfy $x_1 \in \#\Sigma^* \setminus \Sigma^+\#\Sigma^*$ and $x_2 \in \Sigma^*\# \setminus \Sigma^*\#\Sigma^+$ or *vice versa* (to accept this, note that the two maximal proper infixes from Γ^* of a word $a_1a_2\dots a_n$ with $a_i \in \Gamma$ are $a_2a_3\dots a_n$ and $a_1a_2\dots a_{n-1}$). Hence the set $\#\{a,b\}^*\#$ is definable in $\mathcal{M}$ by

$$w \in \{a,b,\#\}^* \wedge \\ \exists x_1, x_2 \left(\begin{array}{l} \text{neighbor}(x_1,w) \wedge \text{neighbor}(x_2,w) \\ \wedge x_1 \in \#\Sigma^* \setminus \Sigma^+\#\Sigma^* \wedge x_2 \in \Sigma^*\# \setminus \Sigma^*\#\Sigma^+ \end{array} \right).$$

Above, we saw that $w \in \{a,b,\#\}$ as well as neighbor(x_i,w) can be expressed in Σ_2 while $x_1 \in \#\Sigma^* \setminus \Sigma^+\#\Sigma^*$ is definable in Σ_0. Thus, the set $\#\{a,b\}^*\#$ is Σ_2-definable in $\mathcal{M}$.

(I5) Since $\#\{a,b\}^*$ is the intersection of the Σ_2-set $\{a,b,\#\}^*$ and the Σ_0-set $\#\Sigma^* \setminus \Sigma^+\#\Sigma^*$, it is Σ_2-definable in $\mathcal{M}$. Similarly, the set $\{a,b\}^*\#$ is Σ_2.

(I6) Let $x = x_1\#x_2 \in \{a,b\}^*\#\{a,b\}^*$ and $w \in \{a,b\}^*$. Then $w = x_1$ iff the largest infix $y \in \{a,b\}^*\#$ of x is an upper neighbor of w (within Γ^*):

$$\exists y \left(\begin{array}{cl} & y \in \{a,b\}^*\# \wedge \text{neighbor}(w,y) \wedge y \le x \\ \wedge & \forall z((z \in \Sigma^*\# \setminus \Sigma^*\#\Sigma^+ \wedge z \le x) \Rightarrow z \le y) \end{array} \right).$$

The set $\{a,b\}^*\#$ and the relation neighbor are Σ_2-definable, and the set $\Sigma^*\# \setminus \Sigma^*\#\Sigma^+$ is Σ_0-definable. Hence this formula is Σ_2. We can write down an analogous Σ_2-formula that holds iff $w = x_2$.

(I7) Let $u, \overline{u} \in \{a,b\}^*$. Then $ua = \overline{u}$ iff

$$\exists x, x', x'' \left(\begin{array}{cl} & \text{neighbor}(u,x) \wedge \text{neighbor}(x,x') \wedge \text{neighbor}(x',x'') \\ \wedge & x, x' \in \bot\{a,b\}^* \wedge x'' \in \Sigma^*\# \setminus \Sigma^*\#\Sigma^+ \\ \wedge & a\# \le x'' \wedge u < \overline{u} < x'' \end{array} \right).$$

The first line expresses that $u < x < x' < x''$ is a maximal chain from u to x'' within Γ^*, *i.e.*, any element of this sequence results from the previous one by prefixing or suffixing by one element of Γ. Since $u \in \{a,b\}^*$ and $x, x' \in \bot\{a,b\}^*$, we get $x = \bot u$ and $x' = \bot uc$ for some $c \in \{a,b\}$. Since $\#$ is a suffix of x'', we also have $x'' = \bot uc\#$. Now $a\# \le x''$ forces $c = a$. Finally, $u < \overline{u}$ says that $\overline{u}$ is a proper extension of u. Since this extension belongs to $\{a,b\}^*$ and is an infix of $x'' = \bot ua\#$, the formula expresses indeed $ua = \overline{u}$.

Since all the conjuncts in the first two lines are at most Σ_2, the whole formula is Σ_2-expressible.

Lemma 5.4. *Let $\mathcal{I} = (u_i,v_i)_{i\le 7}$ be some instance of Post's correspondence problem with $u_i, v_i \in \{a,b\}^*$. There is a Σ_2-formula $\rho(x,\overline{x})$ such that, for $x, \overline{x} \in \{a,b\}^*\#\{a,b\}^*$, we have $\mathcal{M} \models \rho(x,\overline{x})$ iff $x = u\#v$ and $\overline{x} = uu_i\#vv_i$ for some $i \le 7$ and $u, v \in \Gamma^*$.*

Proof. Let $x, \overline{x} \in \{a,b\}^*\#\{a,b\}^*$ and let $u, v, \overline{u}, \overline{v} \in \{a,b\}^*$. Then, by (I6), we can express by a Σ_2-formula that $x = u\#v$ and $\overline{x} = \overline{u}\#\overline{v}$. Now suppose

$u_1 = a_1 a_2 \dots a_n$ with $a_i \in \{a, b\}$. Then $\overline{u} = uu_1$ iff

$$\exists x_0, x_1, x_2, \dots x_n \left(u = x_0 \wedge \bigwedge_{i=1}^{n} x_i = x_{i-1} a_i \wedge x_n = \overline{u} \right) .$$

By (I7), this is a Σ_2-formula. Hence the following is expressible in Σ_2:

$$\exists u, v, \overline{u}, \overline{v} \begin{pmatrix} & x = u\#v \wedge \overline{x} = \overline{u}\#\overline{v} \\ \wedge & \bigvee_{i=1}^{7} (\overline{u} = uu_i \wedge \overline{v} = vv_i) \end{pmatrix} . \qquad \square$$

(I8) A word $w \in \Sigma^*$ belongs to $\perp\{a, b, \#, \perp\}^* \perp \{a, b, \#, \perp\}^* \perp$ iff there are words x' and x'' in the set $\{a, b, \#, \perp\}^*$ with $\perp \leq x'$, $x'\perp = x''$ and $\perp x'' = w$. To express $x'\perp = x''$, we proceed as in (I7) using $\perp'$ instead of $\perp$ and $\#'$ instead of $\#$. Symmetrically, we can express $\perp x'' = w$ by a Σ_2-formula. Hence membership in this set is expressible in Σ_2.

(I9) A word $w \in \Sigma^*$ belongs to $\perp\{a, b, \#\}^* \perp \{a, b, \#\}^* \perp$ iff it is an element of the language $\perp\{a, b, \#, \perp\}^* \perp \{a, b, \#, \perp\}^* \perp$ and there is no proper infix $x < w$ of w with $x \in \perp\{a, b, \#, \perp\}^* \perp \{a, b, \#, \perp\}^* \perp$. In other words, w belongs to this set iff

$$\begin{aligned} & w \in \perp\{a, b, \#, \perp\}^* \perp \{a, b, \#, \perp\}^* \perp \\ & \wedge \forall x ((x \leq w \wedge x \in \perp\{a, b, \#, \perp\}^* \perp \{a, b, \#, \perp\}^* \perp) \Rightarrow x = w). \end{aligned}$$

Since $w \in \perp\{a, b, \#, \perp\}^* \perp \{a, b, \#, \perp\}^* \perp$ is expressible in Σ_2 and appears in a negative position in the second conjunct, this formula is $B\Sigma_2$.

Proposition 5.5. *Let $\mathcal{I} = (a_i, b_i)_{i \leq 7}$ be some instance of Post's correspondence problem with $a_i, b_i \in \Gamma^*$. There is a Σ_4-sentence φ such that $\mathcal{M} \models \varphi$ iff $\mathcal{I}$ has a solution. Since the formula φ can be constructed effectively from $\mathcal{I}$, the Σ_4-theory of $\mathcal{M}$ is undecidable.*

Proof. Let ρ be the Σ_2-formula from Lemma 5.4. Similarly to (I6), there exist Σ_2-formulas φ_1 and φ_2 such that $x = x_1 \perp x_2 \in \{a, b, \#\}^* \perp \{a, b, \#\}^*$ and $u \in \{a, b, \#\}^*$ satisfy $\varphi_i(x, u)$ iff $u = x_i$ for $i = 1, 2$.

Now let

$$R = \{\perp u \perp v \perp \mid u, v \in \{a, b\}^* \# \{a, b\}^* \text{ and } \mathcal{M} \models \rho(u, v)\}$$

and $x \in \perp\{a, b\}^* \# \{a, b\}^* \perp \{a, b\}^* \# \{a, b\}^* \perp$. Then $x \in R$ iff the following holds

$$\exists x_1, x_2, x_3, y \begin{pmatrix} & x = y\perp \wedge y = \perp x_1 \\ \wedge & \varphi_1(x_1, x_2) \wedge \varphi_2(x_1, x_3) \\ \wedge & \rho(x_2, x_3) \end{pmatrix} .$$

The first line ensures $x = \perp x_1 \perp$. Hence $x_1 = u \perp v$ for some $u, v \in \{a, b\}^* \# \{a, b\}^*$. By the second line, $u = x_2$ and $v = x_3$. Hence, by the fourth line, this formula expresses indeed $x \in R$ (provided $x \in \perp\{a, b\}^* \# \{a, b\}^* \perp \{a, b\}^* \# \{a, b\}^* \perp$).

Note that this formula is Σ_2 by (I7) and Lemma 5.4. Now consider the following statement

$$\exists w \left(\begin{array}{ll} & w \in \perp\{a,b,\perp,\#\}^*\perp \\ \wedge & \forall x(x \leq w \Rightarrow (x \notin \#\{a,b\}^*\# \cup \perp\{a,b\}^*\perp)) \\ \wedge & \forall x((x \leq w \wedge x \in \perp\{a,b,\#\}^*\perp\{a,b,\#\}^*\perp) \Rightarrow x \in R) \\ \wedge & \perp\#\perp \leq w \wedge \exists u(u\#u \leq w \wedge u \in \{a,b\}^+) \end{array} \right).$$

We show that the PCP-instance $\mathcal{I}$ has a solution iff this formula is satisfied in $\mathcal{M}$. The conjunction of the first two lines expresses $x \in (\perp\Gamma^*\#\Gamma^*)^+\perp$, *i.e.*,

$$x = (\perp u_1\#v_1)(\perp u_2\#v_2)\dots(\perp u_n\#v_n)\perp$$

for some $n \geq 1$ and $u_i, v_i \in \{a,b\}^*$. The last line expresses that there are $i, j \leq n$ with $u_i = v_i = \varepsilon$ and $u_j = v_j \neq \varepsilon$. Finally, the implication in the third line expresses that, for any $i < n$, we have $\perp u_i\#v_i\perp u_{i+1}\#v_{i+1}\perp \in R$, *i.e.*, $u_{i+1} = u_i a_k$ and $v_{i+1} = v_i b_k$ for some $k \leq 7$. Therefore, indeed, $\mathcal{I}$ has a solution iff this statement holds.

Next we show that the statement above is expressible by a Σ_4-formula. The first line is equivalent to

$$w \in \Gamma^* \wedge \#', \perp' \not\leq w \wedge \exists y, z(y = \perp x \wedge w = y\perp).$$

Since $w \in \Gamma^*$, $y = \perp x$, and $w = y\perp$ are Σ_2-expressible (*cf.* (I7)), the first line is Σ_2-expressible as well. The second conjunct is Π_3 since the conclusion is Σ_2. Now consider the third line. The condition of the implication is $B\Sigma_2$-expressible (I9) and the conclusion is Σ_2. Hence the third line is a Π_3-formula. The last conjunct is equivalent to

$$\exists y, u(\perp y\perp \leq w \wedge \perp \not\leq y \wedge (y,u) \text{ satisfies the two formulas from (I6)})$$

which is Σ_2. Thus, all the conjuncts are at most Π_3 which places the whole formula in Σ_4. □

Theorem 5.6. *The Σ_4-theory of $(\{\alpha,\beta\}^*, \leq)$ is undecidable.*

Proof. By Proposition 5.5, we have to reduce the Σ_4-theory of the extended structure $\mathcal{M}$ to the Σ_4-theory of $(\Sigma^*, \leq)$. We proceed similarly to the proof of Theorem 2.3. Consider the structure $\mathcal{M}_n = (\Sigma^*, \leq, (w)_{w\in\Sigma^+,|w|\leq n})$ for $n \in \mathbb{N}$. We will successively reduce the Σ_4-theory of $\mathcal{M}_{n+1}$ to that of $\mathcal{M}_n$. By Proposition 5.5, the Σ_4-theory of $\mathcal{M} = \mathcal{M}_{14}$ is undecidable. Hence, by induction, the undecidability of the Σ_4-theory of $\mathcal{M}_0 = (\Sigma^*, \leq)$ will follow.

First assume $n \geq 2$ and let $w \in \Sigma^+$ with $|w| = n+1$.

(1) Suppose $w = \alpha^{n+1}$. Then w is the unique word of length $n+1$ whose only infix of length n is α^n. Hence α^{n+1} is the only word w satisfying the

following Π_1-formula $\varphi_{\alpha^{n+1}}(w)$ in the structure $\mathcal{M}_n$:

$$\alpha^n < w \wedge \forall x(\alpha^n < x \leq w \Rightarrow x = w) \wedge \bigwedge_{v \in \Sigma^n \setminus \{\alpha^n\}} \neg v \leq w \ .$$

Similarly, β^{n+1} is the unique word satisfying some Π_1-formula in $\mathcal{M}_n$.

(2) Suppose there is i such that the letters no. i and $i+2$ in w differ. Let w_1 and w_2 be the two infixes of w of length n (since w does not belong to $\alpha^* \cup \beta^*$, there are two such distinct infixes). Then there are $a, b \in \Sigma$ with $w = aw_1 = w_2b$. Suppose v is another word of length $n+1$ containing both, w_1 and w_2. Then there are $c, d \in \Sigma$ with $w_1c = dw_2$. Since $w = w_2b$, letter no. i in w equals letter no. i in w_2. Since $w_1c = dw_2$, it equals letter no. $i+1$ in w_1. Now $aw_1 = w$ implies that it equals letter no. $i+2$ in w, contradicting our assumption on i and w. Hence w is the only word of length $n+1$ containing both, w_1 and w_2. More formally, it is the only word satisfying the following Π_1-formula φ_w in the structure $\mathcal{M}_n$:

$$w_1, w_2 < w \wedge \forall x(w_1 < x \leq w \Rightarrow x = w).$$

(3) Finally, suppose α^2 and β^2 are no infix of w. Then $w \in \{(\alpha\beta)^m, (\beta\alpha)^m\}$ (with $2m = n+1$ if n is odd) or $w \in \{(\alpha\beta)^m\alpha, (\beta\alpha)^m\beta\}$ (with $2m = n$ if n is even). Suppose w starts with α and let w' be the prefix of length n of w. The following distinguishes w from the other element of the respective set: there is an upper neighbor in $(\Sigma^*, \leq)$ containing $\alpha^2\beta$ as an infix. Thus, $(\alpha\beta)^m$ ($(\alpha\beta)^m\alpha$, resp.) is the only word w satisfying the following Σ_2-formula φ_w:

$$\begin{array}{ll} & \alpha^2, \beta^2 \not\leq w \\ \wedge & w' \leq w \wedge \forall y(w' < y \leq w \Rightarrow y = w) \\ \wedge & \exists y(w, \alpha^2\beta < y \wedge \forall x(w < x \leq y \Rightarrow x = y)). \end{array}$$

Since this formula uses the constant $\alpha^2\beta$, it belongs to the language of the structure $\mathcal{M}_n$ for $n \geq 3$, only. If $n = 2$, it is a formula in the language of the extended structure $(\mathcal{M}_2, \alpha^2\beta)$. If the first letter of w is β, we can derive an analogous formula.

Now let φ be some Σ_4-sentence in the language of $\mathcal{M}_{n+1}$. Then let $\overline{\varphi}$ be the following Σ_4-sentence

$$\exists (x_w)_{w \in \Sigma^{n+1}} \left(\bigwedge_{w \in \Sigma^{n+1}} \varphi_w(x_w) \wedge \varphi' \right)$$

where φ' is obtained from φ by replacing any occurrence of the constant $w \in \Sigma^{n+1}$ by the variable x_w. Provided $n \geq 3$, this is a Σ_4-sentence in the language of $\mathcal{M}_n$ and we have $\mathcal{M}_n \models \overline{\varphi}$ iff $\mathcal{M}_{n+1} \models \varphi$. If $n = 2$, the formulas $\varphi_{\alpha\beta\alpha}$ and $\varphi_{\beta\alpha\beta}$ contain occurrences of the constant $\alpha^2\beta$ and $\beta^2\alpha$, resp. Replacing them by $x_{\alpha^2\beta}$

and $x_{\beta^2\alpha}$, resp., turns $\overline{\varphi}$ into a sentence in the language of $\mathcal{M}_2$ with $\mathcal{M}_2 \models \overline{\varphi}$ iff $\mathcal{M}_3 \models \varphi$.

So far, we reduced the undecidable Σ_4-theory of $\mathcal{M}$ to the Σ_4-theory of $\mathcal{M}_2$ which is therefore undecidable as well. Next, we reduce this latter theory to the Σ_4-theory of the structure $\mathcal{M}' = (\Sigma^*, \leq, \alpha, \beta, \alpha\beta)$. As above, it suffices to produce Π_1-formulas defining $\beta\alpha$, α^2 and β^2 in this structure:

$$\begin{array}{lcl} w = \beta\alpha & \iff & \mathcal{M}' \models w \neq \alpha\beta \wedge \alpha, \beta < w \wedge \forall x(\alpha < x \leq w \Rightarrow x = w) \\ w = \alpha^2 & \iff & \mathcal{M}' \models \beta \not\leq w \wedge \alpha < w \wedge \forall x(\alpha < x \leq w \Rightarrow x = w) \\ w = \beta^2 & \iff & \mathcal{M}' \models \alpha \not\leq w \wedge \beta < w \wedge \forall x(\beta < x \leq w \Rightarrow x = w). \end{array}$$

The word $\alpha\beta$ cannot be defined in the structure $\mathcal{M}_1$ since taking the mirror image of a word is an automorphism of $\mathcal{M}_1$ that maps $\alpha\beta$ to $\beta\alpha$.

For a Σ_4-sentence φ in the language of $\mathcal{M}'$, we consider the formula $\overline{\varphi}$

$$\exists x_{\alpha\beta} : (\varphi' \wedge \forall x(x \leq x_{\alpha\beta} \Leftrightarrow (x = x_{\alpha\beta} \vee x \leq \alpha \vee x \leq \beta)))$$

where φ' results from φ by replacing any occurrence of $\alpha\beta$ by $x_{\alpha\beta}$. The universally quantified statement ensures $x_{\alpha\beta} \in \{\alpha\beta, \beta\alpha\}$. Hence $(\Sigma^*, \leq, \alpha, \beta, \alpha\beta) \cong (\Sigma^*, \leq, \alpha, \beta, x_{\alpha\beta})$ and therefore

$$(\Sigma^*, \leq, \alpha, \beta, \alpha\beta) \models \varphi \iff (\Sigma^*, \leq, \alpha, \beta, x_{\alpha\beta}) \models \varphi'.$$

But this implies $\mathcal{M}' \models \varphi$ iff $\mathcal{M}_1 \models \overline{\varphi}$, *i.e.*, we reduced the undecidable Σ_4-theory of $\mathcal{M}'$ to the Σ_4-theory of $\mathcal{M}_1$.

Finally, let φ be a Σ_4-sentence in the language of $\mathcal{M}_1$. Consider the following Σ_4-sentence $\overline{\varphi}$ in the language of $(\Sigma^*, \leq)$:

$$\exists x_\alpha, x_\beta \left(\begin{array}{ll} & \varphi' \wedge x_\alpha \neq x_\beta \\ \wedge & \exists x_\varepsilon \forall x (x < x_\alpha \iff (x = x_\varepsilon \vee x = x_\alpha) \\ \wedge & \exists x_\varepsilon \forall x (x < x_\beta \iff (x = x_\varepsilon \vee x = x_\beta) \end{array} \right)$$

where φ' is obtained from φ by replacing any occurrence of α or β by x_α or x_β, resp. □

Corollary 5.7. *Let $\Sigma = \{a_1, a_2, \ldots, a_n\}$ be an alphabet with at least two elements. Then the Σ_4-theory of the infix order $(\Sigma^*, \leq)$ is undecidable.*

Proof. This can be proved as Corollary 2.4. □

6. Forests

A *forest* is a finite Σ-labeled partial order $(V, \leq, \lambda)$ where any of the sets $\{x \in V \mid x \leq y\}$ is linearly ordered. For two forests s and t, let $s \hookrightarrow t$ denote the existence of an embedding of s into t (*i.e.*, of an injective mapping $f : V_s \to V_t$ such that $\lambda_s(v) = \lambda_t(f(v))$ and $v \leq w \iff f(v) \leq f(w)$ for any $v, w \in V_s$). By Kruskal's theorem, $\hookrightarrow$ is a well quasi order on the set $\mathcal{F}_\Sigma$ of all forests. Note that any term from $T(\Gamma)$ can be seen as a forest with $\Sigma = \Gamma$. By [3], the positive Σ_1-theory of $(T(\Gamma), \Gamma, \hookrightarrow)$ is decidable. We show that the related structure $(\mathcal{F}_\Sigma, \hookrightarrow)$ has an undecidable theory.

Corollary 6.1. *Let Σ be a finite alphabet with at least two elements. Then the theory of $(\mathcal{F}_\Sigma, \hookrightarrow)$ is undecidable.*

Proof. As before, we identify a word $w \in \{a, b\}^*$ with the labeled linear order of length $|w|$ where i is mapped to the ith letter of w. In this sense, $\{a, b\}^*$ is a subset of $\mathcal{F}_\Sigma$ – we show that it is definable which implies the undecidability by Theorem 2.3.

We identify the letter $a \in \Sigma$ with the singleton tree whose only node is labeled a. The empty forest is the only one that does not embed any other forest. A forest t is of size at most n iff there is no sequence of distinct forests $t_0 \hookrightarrow t_1 \hookrightarrow \ldots \hookrightarrow t_n \hookrightarrow t$. Thus, for any n, the set of forests of size at most n is definable. For $a, b \in \Sigma$ let an ab-forest be a forest of size 2 that embeds a, b, and no further forests of size 1. For $a, b \in \Sigma$, let t_{ab} be the disjoint union of the trees a and b. Further, ℓ_{ab} is the 2-elements linear order whose minimal node is labeled a and the other one carries b. Then t_{bb} and ℓ_{bb} are the only bb-forests, and t_{ab}, ℓ_{ab}, and ℓ_{ba} are the only ab-forests for $a \neq b$. We want to distinguish t_{ab} from the other ab-forests. First, let $a \neq b$ and let t be any ab-forest. We consider the set $M(t)$ of forests t' of size 3 such that no bb-forest embeds into t' and t is the only ab-forest that embeds into t'. Then $M(t_{ab})$ contains just 2 elements while $M(\ell_{ab})$ is a three-elements set. Hence t_{ab} as well as the set $\{\ell_{ab}, \ell_{ba}\}$ are definable for $a \neq b$. Then ℓ_{bb} is the unique bb-forest that embeds into some forest t of size 3 with $\ell_{ab}, \ell_{ba} \hookrightarrow t$. Hence, ℓ_{bb} and therefore t_{bb} are definable.

Hence $\{a, b\}^*$ is the set of forests t satisfying

$$\bigwedge_{c \in \Sigma \setminus \{a,b\}} c \not\hookrightarrow t \wedge t_{ab}, t_{aa}, t_{bb} \not\hookrightarrow t \ .$$

Thus, the theory of $(\{a, b\}^*, \hookrightarrow)$ can be interpreted in the theory of $(\mathcal{F}_\Sigma, \hookrightarrow)$ which is therefore undecidable by Theorem 2.3. □

7. Open questions

For some partial orders on the set of words, we showed the undecidability of small fragments of the first-order theory. In this context, the following cases are open

(1) The Σ_2-theory of $(\Sigma^*, \hookrightarrow)$ for $|\Sigma| \geq 2$.
(2) The Σ_n-theory of $(\Sigma^*, \precsim)$ for $2 \leq n \leq 4$ and $|\Sigma| \geq 4$. Nothing is known for $|\Sigma| = 3$ beyond the Σ_1-theory, smaller alphabets have decidable full theories.
(3) The Σ_n-theory of $(\Sigma^*, \leq)$ for $2 \leq n \leq 3$ and $|\Sigma| \geq 2$.

I would have liked some results on the homomorphism preorder of forests [24] since this question by V. Selivanov initiated the research reported in this paper.

References

[1] F. Baader and T. Nipkow, *Term Rewriting and All That.* Cambridge University Press (1998).

[2] A. Björner, The Möbius function of subword order, in *Invariant theory and tableaux*, IMA Springer. *Math. Appl.* **19** (1990) 118–124.

[3] A. Boudet and H. Comon, About the theory of tree embedding, in *TAPSOFT'93*, Springer. *Lect. Notes Comp. Sci.* **668** (1993) 376–390.

[4] J.R. Büchi, Transfinite automata recursions and weak second order theory of ordinals, in *Logic, Methodology and Philos. Sci.*, North Holland, Amsterdam (1965) 3–23.

[5] J.R. Büchi and S. Senger, Coding in the existential theory of concatentation. *Archiv Math. Logik* **26** (1986) 101–106.

[6] H. Comon and R. Treinen, Ordering constraints on trees, in *CAAP'94*, Springer. *Lect. Notes Comp. Sci.* **787** (1994) 1–14.

[7] H. Comon and R. Treinen, The first-order theory of lexicographic path orderings is undecidable. *Theor. Comput. Sci.* **176** (1997) 67–87.

[8] D.E. Daykin, To find all "suitable" orders of $0,1$-vectors. *Congr. Numer.* **113** (1996) 55–60. Festschrift for C. St. J. A. Nash-Williams.

[9] N. Dershowitz, Orderings for term rewriting systems. *Theor. Comput. Sci.* **17** (1982) 279–301.

[10] V.G. Durnev, Undecidability of the positive $\forall\exists^3$-theory of a free semigroup. *Sibirsky Matematicheskie Jurnal* **36** (1995) 1067–1080.

[11] F.D. Farmer, Homotopy spheres in formal languages. *Stud. Appl. Math.* **66** (1982) 171–179.

[12] A. Finkel and Ph. Schnoebelen, Well-structured transition systems everywhere! *Theor. Comput. Sci.* **256** (2001) 63–92.

[13] K. Gödel, Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I. *Monatshefte für Mathematik und Physik* **38** (1931) 173–198.

[14] T. Harju and L. Illie, On quasi orders of words and the confluence property. *Theor. Comput. Sci.* **200** (1998) 205–224.

[15] G. Higman, Ordering by divisibility in abstract algebras. *Proc. London Math. Soc.* **2** (1952) 326–336.

[16] S. Kosub and K. Wagner, The boolean hierarchy of NP-partitions, in *STACS 2000*, Springer. *Lect. Notes Comp. Sci.* **1770** (2000) 157–168.

[17] D. Kuske, Emptiness is decidable for asynchronous cellular machines, in *CONCUR 2000.* Springer. *Lect. Notes Comp. Sci.* **1877** (2000) 536–551.

[18] U. Leck, *Nonexistence of a Kruskal-Katona type theorem for subword orders.* Technical Report 98-06, Universität Rostock, Fachbereich Mathematik (1998).

[19] M. Lothaire, *Combinatorics on Words.* Addison-Wesley (1983).

[20] G.S. Makanin, The problem of solvability of equations in a free semigroup, *Math. Sbornik*, **103** (1977) 147–236. In Russian; English translation in: *Math. USSR Sbornik* **32** (1977).

[21] Y. Matiyasevich, *Hilbert's Tenth Problem.* MIT Press (1993).

[22] P. Narendran and M. Rusinowitch, The theory of total unary rpo is decidable, in *Computational Logic 2000*, Springer. *Lect. Notes Artificial Intelligence* **1861** (2000) 660–672.

[23] M.O. Rabin, Decidability of second-order theories and automata on infinite trees. *Trans. Amer. Math. Soc.* **141** (1969) 1–35.

[24] V.L. Selivanov, Boolean hierarchy of partitions over reducible bases. *Algebra and Logic* **43** (2004) 77–109.

[25] R. Treinen, A new method for undecidability proofs of first order theories. *J. Symbolic Comput.* **14** (1992) 437–458.

Communicated by V. Diekert.
Received May 26, 2004. Accepted February 3, 2005.